

Fast Mixing for Discrete Point Processes*

Patrick Rebeschini

PATRICK.REBESCHINI@YALE.EDU

Amin Karbasi

AMIN.KARBASI@YALE.EDU

Yale Institute for Network Science, 17 Hillhouse Avenue, Yale University, New Haven, CT, 06511, USA

Abstract

We investigate the systematic mechanism for designing fast mixing Markov chain Monte Carlo algorithms to sample from discrete point processes under the Dobrushin uniqueness condition for Gibbs measures. Discrete point processes are defined as probability distributions $\mu(S) \propto \exp(\beta f(S))$ over all subsets $S \in 2^V$ of a finite set V through a bounded set function $f : 2^V \rightarrow \mathbb{R}$ and a parameter $\beta > 0$. A subclass of discrete point processes characterized by submodular functions (which include log-submodular distributions, submodular point processes, and determinantal point processes) has recently gained a lot of interest in machine learning and shown to be effective for modeling diversity and coverage. We show that if the set function (not necessarily submodular) displays a natural notion of decay of correlation, then, for β small enough, it is possible to design fast mixing Markov chain Monte Carlo methods that yield error bounds on marginal approximations that do not depend on the size of the set V . The sufficient conditions that we derive involve a control on the (discrete) Hessian of set functions, a quantity that has not been previously considered in the literature. We specialize our results for submodular functions, and we discuss canonical examples where the Hessian can be easily controlled.

Keywords: Discrete point processes, MCMC, fast mixing, submodular functions, decay of correlation, Hessian of set functions

1. Introduction

Probabilistic modeling and inference techniques have become essential tools for analyzing data and making predictions in a variety of real-world settings. Graphical models (Wainwright and Jordan, 2008) have provided an appealing framework to express dependencies among variables through a graph structure. A broad class of such models that have been widely used in machine learning is represented by Markov random fields, where the probability distribution of a collection of n random variables $X := (X^1, \dots, X^n)$ is defined as a product of non-negative potentials ϕ_C over maximal cliques $C \subseteq G$ of an undirected graph $G = (V, E)$, i.e., $\mu(x) := \mathbf{P}(X = x) = 1/Z \prod_C \phi_C(x^C)$, where we adopted the notation $x^C := \{x^i\}_{i \in C}$. Here, Z is the normalization factor, often called the *partition function*, and it is known to be hard to compute exactly (Jerrum and Sinclair, 1993), or even to approximate (Goldberg and Jerrum, 2007). Perhaps the most prominent example of Markov networks, with many applications in machine learning, is the pairwise Markov random field, also called Ising model, where cliques are defined on edges between pairs of variables.

These examples can be seen as instances of a general class of probabilistic models that we refer to as *discrete point processes*, which are defined as

$$\mu(S) = \frac{1}{Z} \exp(\beta f(S)) \quad \text{for } S \subseteq V, \quad (1)$$

* This is the full version of a paper in the 28th Annual Conference on Learning Theory (COLT), 2015.

where $f : 2^V \rightarrow \mathbb{R}$ is a bounded set function, $Z := \sum_{S \subseteq V} \exp(\beta f(S))$ is the partition function, and β is a strictly positive real constant that parametrizes the distribution. Discrete point processes have been widely studied in mathematics and statistical physics (Daley and Vere-Jones, 2007)[ch. 5], where they have been traditionally used to model particle processes and neural spiking activity, among many others. Note that distributions over subsets of V are isomorphic to distributions of $n := |V|$ binary random variables $X^1, \dots, X^n \in \{0, 1\}$.

A subclass of discrete point processes, referred to as log-submodular (log-supermodular) distributions, has recently been investigated in Djolonga and Krause (2014), where the set function f is taken to be *submodular* (respectively, *supermodular*), i.e., characterized by the property that the difference in the value of the function when an element is added to a set, the so-called *marginal gain*, decreases (increases) as the cardinality of the set increases. Throughout this paper, the discrete derivative, or marginal gain, of f is defined as $\Delta_i f(S) := f(S \cup \{i\}) - f(S)$. The function f is submodular (supermodular) if for any $S \subseteq S' \subseteq V$ and any $i \in V \setminus S'$, it holds $\Delta_i f(S) \geq \Delta_i f(S')$ ($\Delta_i f(S) \leq \Delta_i f(S')$). Under some regularity conditions (discussed in Djolonga and Krause (2014)), pairwise Markov random fields are also a special case of log-submodular point processes, as are determinantal point processes (Kulesza and Taskar, 2012), (Hough et al., 2006). Here, the submodular function is defined as $f(S) = \log \det(L_S)$ where $L \in \mathbb{R}^{n \times n}$ is a positive definite matrix and L_S is the square submatrix of L that is indexed by S . Determinantal point processes have been extensively used in physics and machine learning to model negative correlations, giving rise, for instance, to diverse sets of items in recommendations. Another related subclass of discrete point processes, called submodular (supermodular) point processes, has been recently proposed in Iyer and Bilmes (2015), where $\mu(S) \propto f(S)$ with f a non-negative submodular (supermodular) function.

The diminishing return property that characterizes submodularity — which makes submodular functions suitable for applications in several fields, ranging from economics to machine learning — has been extensively investigated in the domain of optimization (see Krause and Golovin (2014) for instance), but its role has yet to be established in the realm of probabilistic inference. Iyer and Bilmes (2015) showed that, in general, computing the partition function in log-submodular distributions and submodular point processes requires exponential complexity. Djolonga and Krause (2014) and Iyer and Bilmes (2015) resort to variational approaches to approximate the partition function. In Djolonga and Krause (2014) the authors provide upper and lower bounds based on sub- and super- gradients (Iyer et al., 2013), showing that the log partition function $\log(Z)$ can be approximated within $O(n)$. In their setting, however, this implies that the error bounds that can be derived from their theory to approximate marginals of the type $\mu(\{S \subseteq V : S \ni i\})$, for a given $i \in V$, deteriorate exponentially with the size of the model $n = |V|$, as can be deduced from their experimental results. In addition, the bounds that they consider depend on the *curvature* $c(f)$ of the submodular function f , a quantity between 0 and 1 that characterizes the deviation from modularity (f is said to be *modular* if $\Delta_i f(S) = \Delta_i f(S')$ for all $S, S' \subseteq V, i \notin S, S'$, and in this case $c(f) = 0$). However, there are trivial examples with very little interactions between random variables (Section 4.1) for which the inference problem has a straightforward solution but $c(f) = 1$, so that the corresponding (lower/upper) bounds on the partition function in Djolonga and Krause (2014) are unbounded and no useful inference can be deduced. In contrast to Markov random fields, whose partition function is typically intractable and hard to approximate, determinantal point processes admit an exact algorithm for marginalization, albeit in time cubic in the size n . In order to avoid this cost, recently Kang (2013) considered a Markov chain Monte Carlo (MCMC) algorithm to sample from determinant point processes. The author claims that this algorithm is fast mixing

(i.e., the Markov chain gets arbitrarily close to equilibrium after a small, $O(n \log n)$, number of steps) with no additional assumptions on the model. However, the proofs of the results in this paper are wrong as they rely on ill-defined couplings between Markov chains.

In this paper we address the inference problem of approximating marginals of the type $\mu(\{S \subseteq V : S \ni i\})$ for discrete point processes defined in Eq. (1) through MCMC methods. In Section 2 we define the general class of local-update algorithms that we consider, and we present the main theoretical result for fast mixing, i.e., Theorem 1, which relies on the Dobrushin uniqueness condition for Gibbs measures (Dobrušin, 1970), (Georgii, 2011). In Section 3 we analyze two specific algorithms within this class, Gibbs sampling and Metropolis-Hastings, and we show that if the set function f satisfies some natural notion of *decay of correlation*, then these algorithms are fast mixing and yield size-free error bounds on marginal approximations. The decay of correlation property that we exploit concerns the decay of the absolute value of the *difference* of marginal gains evaluated at sets differing only by a single element j , as a function of the element i being added (in fact, here the role of i and j is symmetric). More precisely, this property is related to the second order derivatives $\Delta_j \Delta_i f(S) = \Delta_i f(S \cup \{j\}) - \Delta_i f(S) \equiv \Delta_i \Delta_j f(S)$, i.e., to the (discrete) *Hessian* of the function f . If we define $M_{ij} := \max_{S \subseteq V : S \not\ni i, j} |\Delta_j \Delta_i f(S)|$ for each $i \neq j$, and $M_{ii} := 0$ for each i , we show that the Gibbs sampler is fast mixing if the following condition holds

$$\alpha(\beta) \beta \|M\|_\infty \leq \gamma < 1,$$

where $\|M\|_\infty := \max_{i \in V} \sum_{j \in V} M_{ij}$, $\alpha(\beta) := \max_{i \in V} \max_{S \subseteq V \setminus \{i\}} e^{-\beta \Delta_i f(S)}$, and $\gamma \geq 0$ is a quantity that does not depend on n . If the set function is submodular, i.e., the distribution is log-submodular, then we can simplify this condition (Lemma 8). To the best of our knowledge, these results are the first to emphasize the importance of the Hessian of set functions, a quantity that has not been previously investigated even in the optimization domain. Finally, in Section 4 we specialize our results for a number of canonical examples of submodular functions (facility location, cut function, log determinant functions leading to determinantal point processes, and decomposable functions). These examples attest that our general criterion, which a priori involves a combinatorial optimization to compute each term M_{ij} , can often be reduced to a simple-to-check condition. Proofs are given in Appendix A (theory) and Appendix B (applications). As a final remark, we should highlight that submodularity (supermodularity), which is equivalent to $\Delta_j \Delta_i f(S) \leq 0$ ($\Delta_j \Delta_i f(S) \geq 0$) for any $i, j \in V$, $i \neq j$, $S \subseteq V$, $S \not\ni i, j$, is not sufficient to guarantee fast mixing, as displayed by the different convergence behaviors of the Glauber dynamics for Ising models with respect to different values of the inverse temperature β . See Mossel and Sly (2013) and references therein.

Notation. In this paper we adopt the usual vector/matrix notation for distributions, kernels, and functions defined on finite sets. Given two finite sets \mathbb{X} and \mathbb{Y} with respective cardinality $|\mathbb{X}|$ and $|\mathbb{Y}|$, we interpret a probability distribution ρ on \mathbb{X} as $|\mathbb{X}|$ -dimensional row vector, a kernel T from \mathbb{X} to \mathbb{Y} as a matrix in $[0, 1]^{|\mathbb{X}| \times |\mathbb{Y}|}$, and a function $h : \mathbb{Y} \rightarrow \mathbb{R}$ as a $|\mathbb{Y}|$ -dimensional column vector. Hence, we write $\rho T h := \sum_{x \in \mathbb{X}, y \in \mathbb{Y}} \rho(x) T(x, y) h(y)$. For each $x \in \mathbb{X}$, we write T_x to indicate the probability distribution $T_x : y \in \mathbb{Y} \rightarrow T_x(y) := T(x, y)$, and for $m \geq 0$ we write T^m to indicate the m -th power of the matrix T . Given $A \subseteq \mathbb{Y}$, we define the indicator function 1_A as $1_A(y) := 1$ if $y \in A$, $1_A(y) := 0$ if $y \notin A$. Clearly, $\rho(A) = \rho 1_A$. For $y \in \mathbb{Y}$, we will also use the notation 1_y to mean $1_{\{y\}}$. If V is a finite set, given $x = (x^i)_{i \in V}$, we write $x^S := (x^i)_{i \in S}$, for $S \subseteq V$. If $x, y \in \mathbb{R}$, we use $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$.

2. Fast mixing MCMC algorithms for discrete point processes

Throughout this paper, let V be a finite set with cardinality $n := |V|$. Let $\mathbb{S}_0 := \{0, 1\}$ and define $\mathbb{S} := \times_{i \in V} \mathbb{S}_0 = \{0, 1\}^n$. Given $\beta > 0$, we consider the following probability distribution μ on \mathbb{S}

$$\mu(x) := \frac{e^{\beta f(x)}}{\sum_{x \in \mathbb{S}} e^{\beta f(x)}}, \quad (2)$$

for $x \in \mathbb{S}$, where $f : \mathbb{S} \rightarrow \mathbb{R}$ is a given bounded function. In the following we will also consider the isomorphic description of the function f given by the set function $f^* : 2^V \rightarrow \mathbb{R}$ defined as follows, for each $S \subseteq V$: $f^*(S) := f(x(S))$, where $x(S) \in \mathbb{S}$ is defined as $x(S)^i = 1$ if $i \in S$, $x(S)^i = 0$ if $i \notin S$. With an overload of notation, henceforth we refer to f^* as f , leaving to the context the determination of what is meant. In this paper we address the problem of probabilistic inference for discrete point processes defined in (2), that is, the problem of computing guaranteed approximations to marginals probabilities of the type

$$\mu(\{x \in \mathbb{S} : x^i = 1 \forall i \in S\}) = \sum_{x \in \mathbb{S} : x^i = 1, i \in S} \mu(x), \quad (3)$$

for a given $S \subseteq V$. Note that, in general, computing (3) exactly is hard, as computing the normalization function in (2) is $\#P$ -complete, see [Jerrum and Sinclair \(1993\)](#). Hence, we need to resort to approximation schemes. Our goal is to investigate the properties of the function f that make it possible to design time-uniform Markov chains that quickly converge to μ . More specifically, we want to design a transition kernel T on \mathbb{S} such that $\mu T = \mu$, and so that, for any distribution ρ on \mathbb{S} and any function $h : \mathbb{S} \rightarrow \mathbb{R}$, $\rho T^m h$ converges exponentially fast to μh , as m increases. As $|\mathbb{S}| = 2^n$, a general transition kernel T on \mathbb{S} is a matrix with at most $2^n(2^n - 1)$ degrees of freedom (recall that for each $x \in \mathbb{S}$ it has to hold $\sum_{z \in \mathbb{S}} T(x, z) = 1$). In particular, for each $x \in \mathbb{S}$, T_x is a distribution with 2^{n-1} degrees of freedom, which can be difficult to sample from if n is large. To avoid this exponential burden with the cardinality of \mathbb{S} , we restrict our attention on Markov chains that are defined as combinations of *local-update* probability kernels $K^{[i]}$ from \mathbb{S} to \mathbb{S}_0 that we assume we can easily sample from (in particular, this implies that each $K^{[i]}$ should not depend on the normalization function in (2)). Hence, for each $i \in V$ define the transition kernel $T^{[i]}$ on \mathbb{S} as

$$T^{[i]}(x, z) := K^{[i]}(x, z^i)1_{x^{V \setminus \{i\}}}(z^{V \setminus \{i\}}), \quad (4)$$

where, for each $i \in V$, $K^{[i]}$ is a probability kernel from \mathbb{S} to \mathbb{S}_0 so that $T^{[i]}$ leaves μ invariant, namely, $\mu T^{[i]} = \mu$. Label the elements of V as $V = \{i_1, \dots, i_n\}$. We consider the two chains:

$$T_s := T^{[i_1]} \dots T^{[i_n]}, \quad (\text{systematic scan}) \quad (5)$$

$$T_r := \left(\frac{1}{n} \sum_{i \in V} T^{[i]} \right)^n. \quad (\text{random scan}) \quad (6)$$

Markov chains described by these type of transition kernels are usually referred to as *systematic scan* (5) and *random scan* (6) Markov chain Monte Carlo (MCMC)¹. Respectively, they give rise to Algorithm 1 and Algorithm 2, when the chains are run for $m \geq 1$ sweeps with initial distribution ρ .

1. Typically in the literature (see [Dyer et al. \(2009\)](#), for instance) the random scan MCMC sampler is defined as $S := \frac{1}{n} \sum_{i \in V} T^{[i]}$, instead of $T_r = S^n$ as in (6). Our choice in the present context is motivated by the fact that we want to compare random and systematic scan. Note, in fact, that a single application of the kernel T_s in (5) involves updating all coordinates i_1, \dots, i_n , while a single application of S involves updating (uniformly at random) only one coordinate $i \in V$. This is why the right scale to make the comparison involves n iterations of S .

Algorithm 1 Systematic scan MCMC sampler

Sample $X \in \mathbb{S}$ from the distribution ρ ;
for $k = 1, \dots, m$ **do**
 | **for** $j = 1, \dots, n$ **do**
 | | Draw Z^{ij} from the distribution $K_x^{[ij]}$;
 | | Set $X^{ij} \leftarrow Z^{ij}$;
 | **end**
end
Output: $X = (X^i)_{i \in V}$ that is distributed according to ρT_s^m .

Algorithm 2 Random scan MCMC sampler

Sample $X \in \mathbb{S}$ from the distribution ρ ;
for $k = 1, \dots, nm$ **do**
 | Sample $i \in V$ uniformly;
 | Draw Z^i from the distribution $K_x^{[i]}$;
 | Set $X^i \leftarrow Z^i$;
end
Output: $X = (X^i)_{i \in V}$ that is distributed according to ρT_r^m .

The following theorem describes the convergence behavior of the MCMC algorithms T_s and T_r under the so-called *Dobrushin uniqueness condition*. Since the seminal work in Dobrušin (1970) several authors have presented different approaches to establish convergence bounds of the type in Theorem 1. We refer to Dyer et al. (2009) and references therein for a review of results that address fast mixing within the Dobrushin uniqueness framework. Theorem 1 represents the building block for the theory that will be developed in the next sections.

Theorem 1 (Local-update MCMC algorithms for discrete point processes) *For each $i, j \in V$, define the Dobrushin coefficients as:*

$$C_{ij} := \max_{x \in \mathbb{S}} |K_{x^{V \setminus \{j\}} 0^j}^{[i]}(\{0\}) - K_{x^{V \setminus \{j\}} 1^j}^{[i]}(\{0\})|. \quad (7)$$

Let $R \in \mathbb{R}^{n \times n}$ such that $C \leq R$, element-wise. Assume that the Dobrushin uniqueness condition holds, namely,

$$\|R\|_\infty := \max_{i \in V} \sum_{j \in V} R_{ij} \leq \gamma < 1. \quad (8)$$

Then, for any distribution ρ on \mathbb{S} , any natural number $m \geq 0$, and any function $h : \mathbb{S} \rightarrow \mathbb{R}$, we have

$$|\rho T_s^m h - \mu h| \leq \gamma^m \sum_{i \in V} \delta_i(h), \quad |\rho T_r^m h - \mu h| \leq \lambda^m \sum_{i \in V} \delta_i(h), \quad (9)$$

with $\lambda := e^{\gamma-1} < 1$ and $\delta_i(h) := \max_{x \in \mathbb{S}} |h(x^{V \setminus \{i\}} 0^i) - h(x^{V \setminus \{i\}} 1^i)|$.

Remark 2 (On fast mixing) If γ in the Dobrushin uniqueness condition (8) does not depend on n , then the results in Theorem 1 imply that the Markov chains defined by the transition kernels T_s and T_r are fast mixing. Recall that for each $\varepsilon > 0$ the mixing time $\tau(\varepsilon)$ of a Markov chain with transition kernel T and unique invariant distribution μ is defined as (Levin et al., 2008)

$$\tau(\varepsilon) := \min \left\{ m \geq 0 : \sup_{\rho \text{ distribution on } \mathbb{S}} \|\rho T^m - \mu\|_{TV} \leq \varepsilon \right\},$$

where $\|\rho - \rho'\|_{TV} := \max_{A \subseteq \mathbb{S}} |\rho(A) - \rho'(A)|$ is the total variation distance between distributions ρ and ρ' on \mathbb{S} . Clearly, from (9) it follows that $|\rho T_s^m(A) - \mu(A)| \leq n\gamma^m$ and $|\rho T_r^m(A) - \mu(A)| \leq n\lambda^m$, for each $A \subseteq \mathbb{S}$. Hence, we can easily derive the following upper bounds for the mixing time of the systematic scan and random scan Markov chains, respectively,

$$\tau_s(\varepsilon) \leq \left\lceil \frac{\log(n\varepsilon^{-1})}{1 - \gamma} \right\rceil, \quad \tau_r(\varepsilon) \leq \left\lceil \frac{\log(n\varepsilon^{-1})}{1 - \lambda} \right\rceil,$$

which show that the Markov chains are fast mixing, that is, their mixing time is upper bounded by a quantity that scales only logarithmically with the size n of the set V .² In fact, taking the case of the systematic scan Markov chain, for instance, we have

$$\tau(\varepsilon) \leq \min \{m \geq 0 : n\gamma^m \leq \varepsilon\} = \min \left\{ m \geq 0 : m \geq \frac{\log(n\varepsilon^{-1})}{-\log \gamma} \right\} \leq \left\lceil \frac{\log(n\varepsilon^{-1})}{1 - \gamma} \right\rceil,$$

where in the last inequality we used that $\log x \leq x - 1$ for each $x \in \mathbb{R}$.

Remark 3 (On probabilistic inference) If γ in the Dobrushin uniqueness condition (8) does not depend on n , then Theorem 1 yields that the inference problem of approximating marginals of the type (3) can be efficiently addressed via MCMC methods. In fact, choosing $h = 1_A$, with $A = \{x \in \mathbb{S} : x^i = 1, i \in S\}$ for a certain $S \subseteq V$, Theorem 1 provides the following exponentially decreasing error bounds, for any $m \geq 0$ and any distribution ρ on \mathbb{S} :

$$|\rho T_s^m(A) - \mu(A)| \leq |S|\gamma^m, \quad |\rho T_r^m(A) - \mu(A)| \leq |S|\lambda^m,$$

which do not depend on n . As a consequence, if X_1, \dots, X_N are independent random variables distributed according to ρT_s^m , generated as prescribed in Algorithm 1 (analogous results follow for random scan), then $\hat{\mu}_A := \frac{1}{N} \sum_{k=1}^N 1_A(X_k)$ is a biased estimator of $\mu(A)$ with the typical Monte Carlo mean square error bias/variance decomposition:

$$\mathbf{E}[(\hat{\mu}_A - \mu(A))^2] = \underbrace{(\mathbf{E}[\hat{\mu}_A] - \mu(A))^2}_{\text{bias}^2} + \underbrace{\mathbf{E}[(\hat{\mu}_A - \mathbf{E}[\hat{\mu}_A])^2]}_{\text{variance}} \leq (|S|\gamma^m)^2 + \frac{1}{N}.$$

We stress once again that under the current assumptions this error bound does not depend on the set size n . This is in sharp contrast to the upper bounds for marginals produced by the theory developed for the variational methods in Djalonga and Krause (2014); these bounds, in fact, deteriorate exponentially with n , as it can be deduced by the very experimental results presented by the authors.

2. Typically, fast mixing is defined when the mixing time is $O(n \log n)$, not $O(\log n)$. The difference in our setting is due to the definitions of the chains T_s and T_r , which involve a full sweep over n variables.

Remark 4 (On block updates) *Using the results in [Rebeschini and van Handel \(2014\)](#) it is possible to generalize Theorem 1 for MCMC algorithms with blocks updates, in the spirit of [Weitz \(2005\)](#), that is, when the transition kernels $T^{[i]}$, $i \in V$, in (4) are replaced by*

$$T^{[S]}(x, z) := K^{[S]}(x, z^S)1_{x^{V \setminus S}}(z^{V \setminus S}),$$

for each block $S \subseteq V$ in a given family \mathcal{S} satisfying $\bigcup_{S \in \mathcal{S}} S = V$ (for instance, $\mathcal{S} = \{S \subseteq V : |S| = k\}$, for some $k \geq 1$), where $K^{[S]}$ is a probability kernel from \mathbb{S} to $\times_{i \in S} \mathbb{S}_0 = \{0, 1\}^{|S|}$. In this case, for instance, a single step of the block-update random scan Gibbs sampler would read $\frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} T^{[S]}$, which generalizes $\frac{1}{n} \sum_{i \in V} T^{[i]}$ in (6). The new version of Theorem 1 would yield weaker sufficient conditions for fast mixing involving higher values of the parameter β , in the spirit of [Dobrushin and Shlosman \(1985\)](#). However, these conditions would be more convoluted and difficult to analyze, and they would involve a control on the “block” derivatives of the form $\Delta_J \Delta_i f(S)$, with $J \subseteq V$. As the focus of the present work lies on investigating the basic properties of the set function f that makes it possible to efficiently address the inference problem in (2) via MCMC methods, we limit our analysis to local sampling schemes with single site updates.

3. Condition for fast mixing and decay of correlations

In this section we introduce two of the most popular MCMC algorithms that are used in the literature within the local-update framework described by (4): Gibbs sampler and Metropolis-Hastings ([Asmussen and Glynn, 2007](#)). While many variants of these algorithms can also be considered, we restrict our analysis to their most basic implementations, as our goal is to investigate the fundamental principles behind fast mixing for discrete point processes. For both of these algorithms we compute the Dobrushin coefficients (7), and we provide a comparison between them in Lemma 7 below. Then, we apply Theorem 1 to the analysis of the Gibbs sampler algorithm, and we present sufficient conditions for fast mixing, particularly in the case when f is submodular, Lemma 8 below. Henceforth, for each $x \in \mathbb{S}$ let $S(x) \subseteq V$ be defined as follows: $i \in S$ if $x(S)^i = 1$, $i \notin S$ if $x(S)^i = 0$.

Lemma 5 (Gibbs sampler) *For each $i \in V$, in the transition kernel $T^{[i]}$ in (4) choose*

$$K^{[i]}(x, z^i) = \mathbf{E}[X^i = z^i | X^{V \setminus \{i\}} = x^{V \setminus \{i\}}] = \begin{cases} \frac{1}{1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i\}} 0^i))}} & \text{if } z^i = 0, \\ \frac{e^{\beta \Delta_i f(S(x^{V \setminus \{i\}} 0^i))}}{1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i\}} 0^i))}} & \text{if } z^i = 1, \end{cases} \quad (10)$$

where the random variable X has distribution μ . Then, for each $i \in V$ we have $\mu T^{[i]} = \mu$, and the Dobrushin coefficients (7) read, for each $i, j \in V$,

$$C_{ij} := \begin{cases} 0 & \text{if } i = j, \\ \max_{S \subseteq V: S \not\ni i, j} \frac{|e^{\beta \Delta_i f(S)} - e^{\beta \Delta_i f(S \cup \{j\})}|}{(1 + e^{\beta \Delta_i f(S)})(1 + e^{\beta \Delta_i f(S \cup \{j\})})} & \text{if } i \neq j. \end{cases}$$

Lemma 6 (Metropolis-Hastings) *For each $i \in V$, in the transition kernel $T^{[i]}$ in (4) choose*

$$K^{[i]}(x, z^i) = \begin{cases} \alpha^i(x) & \text{if } z^i \neq x^i, \\ 1 - \alpha^i(x) & \text{if } z^i = x^i, \end{cases} \quad \alpha^i(x) := \begin{cases} 1 \wedge e^{-\beta \Delta_i f(S(x^{V \setminus \{i\}} 0^i))} & \text{if } x^i = 1, \\ 1 \wedge e^{\beta \Delta_i f(S(x^{V \setminus \{i\}} 0^i))} & \text{if } x^i = 0. \end{cases} \quad (11)$$

Then, for each $i \in V$ we have $\mu T^{[i]} = \mu$, and the Dobrushin coefficients (7) read, for each $i, j \in V$,

$$\tilde{C}_{ij} := \begin{cases} \max_{S \subseteq V: S \not\ni i} e^{-\beta|\Delta_i f(S)|} & \text{if } i = j, \\ \max_{S \subseteq V: S \not\ni i, j} \left| 1 \wedge e^{\beta \Delta_i f(S)} - 1 \wedge e^{\beta \Delta_i f(S \cup \{j\})} \right| \vee \left| 1 \wedge e^{-\beta \Delta_i f(S)} - 1 \wedge e^{-\beta \Delta_i f(S \cup \{j\})} \right| & \text{if } i \neq j. \end{cases}$$

The implementation of the Gibbs sampler (10) and Metropolis-Hastings (11) is given, respectively, in Algorithm 3 and Algorithm 4 in Appendix A (we only present the random scan versions from Algorithm 2; the systematic scan versions can be obtained analogously from Algorithm 1).

The following Lemma compares the Dobrushin coefficients for the Gibbs sampler, Lemma 5, and the Metropolis-Hastings Markov chain, Lemma 6.

Lemma 7 (Comparison of Dobrushin coefficients) *For each $i, j \in V, i \neq j$, we have*

$$0 = C_{ii} \leq \tilde{C}_{ii}, \quad \frac{1}{4} \tilde{C}_{ij} \leq C_{ij} \leq \tilde{C}_{ij}.$$

From Lemma 7 it follows that $C \leq \tilde{C}$, element-wise. In this respect, it might seem a good idea to investigate conditions for fast mixing for the Metropolis Hastings algorithm, i.e., to upper bound \tilde{C} , as these conditions will immediately yield fast mixing for the Gibbs sampler as well. However, this approach has a main drawback. In fact, while for each $i \in V$ it holds that $C_{ii} = 0$, we have that $\tilde{C}_{ii} \geq 0$, and for a large system (i.e., for a large value of $n = |V|$) we typically expect this quantity to be very close to 1, which would create a problem to establish the Dobrushin uniqueness condition (8). For instance, take the case where the function f is *monotone* (i.e., $\Delta_i f(S) \geq 0$ for each $i \in V$, $S \subseteq V$) and submodular. Then,

$$\tilde{C}_{ii} = \max_{S \subseteq V: S \not\ni i} e^{-\beta|\Delta_i f(S)|} = \max_{S \subseteq V: S \not\ni i} e^{-\beta \Delta_i f(S)} = e^{-\beta \Delta_i f(V \setminus \{i\})}.$$

Typically, unless f is modular or “close” to modular (in the sense that its *curvature* is close to zero, see Section 4.1), if n is large we expect the marginal gain of f when an element $i \in V$ is added to $V \setminus \{i\}$ to be close to 0, that is, $\Delta_i f(V \setminus \{i\}) \approx 0$, which yields $\tilde{C}_{ii} \approx 1$ if $\beta \approx 1$. This issue is not present in the Gibbs sampler, as $C_{ii} = 0$ for each $i \in V$. For this reason, henceforth we focus on the Gibbs sampler and we provide sufficient conditions for it to be fast mixing.

Lemma 8 (Fast mixing Gibbs samplers for discrete point processes) *Assume that the following condition holds:*

$$\alpha(\beta) \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left| 1 - e^{\beta \Delta_j \Delta_i f(S)} \right| \leq \gamma < 1, \quad (12)$$

where $\alpha(\beta) := \max_{i \in V} \max_{S \subseteq V \setminus \{i\}} e^{-\beta \Delta_i f(S)}$. Let T_s and T_r be the Gibbs samplers defined as in Section 2, with $K^{[i]}$, $i \in V$, defined as in (10). Then, for any distribution ρ on \mathbb{S} , any natural number $m \geq 0$, and any function $h : \mathbb{S} \rightarrow \mathbb{R}$, we have

$$|\rho T_s^m h - \mu h| \leq \gamma^m \sum_{i \in V} \delta_i(h), \quad |\rho T_r^m h - \mu h| \leq \lambda^m \sum_{i \in V} \delta_i(h),$$

where $\lambda := e^{\gamma-1} < 1$ and $\delta_i(h) := \max_{x \in \mathbb{S}} |h(x^{V \setminus \{i\}} 0^i) - h(x^{V \setminus \{i\}} 1^i)|$. If γ does not depend on n , then the chains are fast mixing. In particular, if f is submodular³, then condition (12) reads

$$\alpha(\beta) \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left(1 - e^{\beta \Delta_j \Delta_i f(S)}\right) \leq \gamma < 1, \quad (13)$$

with $\alpha(\beta) = \max_{i \in V} e^{-\beta \Delta_i f(V \setminus \{i\})}$. If f is monotone, clearly $\alpha(\beta) \leq 1$.

Note that using the inequality $1 - e^x \leq -x$ for each $x \in \mathbb{R}$, condition (12) can be replaced by the stronger (but simpler) condition $\alpha(\beta) \beta \|M\|_\infty \leq \gamma < 1$, where $M_{ij} := \max_{S \subseteq V: S \not\ni i, j} |\Delta_j \Delta_i f(S)|$ for each $i \neq j$, and $M_{ii} := 0$ for each i . This condition makes it manifest the property of the function f that renders the inference problem tractable via local-update MCMC methods. As highlighted in Remark 2 and also in Remark 3, in order for the Markov chains to be fast mixing we need γ to be independent of the set size $n = |V|$. In particular, we need $\|M\|_\infty$ to be upper bounded by a constant that does not depend on n . This is achieved, for instance, if (V, d) is a metric space and the function f displays one of the following forms of *decay of correlation* with respect to the metric d .

- (Exponential decay of correlations) Assume that for each $i, j \in V, i \neq j$ we have

$$M_{ij} := \max_{S \subseteq V: S \not\ni i, j} |\Delta_j \Delta_i f(S)| \leq \alpha e^{-\alpha' d(i, j)},$$

where $\alpha, \alpha' > 0$ are two constants so that $\|M\|_\infty \leq \max_{i \in V} \sum_{j \in V} \alpha e^{-\alpha' d(i, j)} \leq c$, where c does not depend on n .

- (Finite-range correlations) Assume that for each $i, j \in V, i \neq j$, there exists $r > 0$ such that

$$M_{ij} := \max_{S \subseteq V: S \not\ni i, j} |\Delta_j \Delta_i f(S)| \leq \begin{cases} c & \text{if } d(i, j) \leq r, \\ 0 & \text{if } d(i, j) > r, \end{cases}$$

where $c \geq 0$ does not depend on n . For each $i \in V$ define $N(i) := \{j \in V : d(i, j) \leq r\}$, and assume that $N := \max_{i \in V} |N(i)|$ does not depend on n . Then, clearly, $\|M\|_\infty \leq cN$.

In the remaining of this paper we assume that the function f is submodular, and we investigate condition (13). In this case, note that the quantity $\alpha(\beta)$ involves an optimization over only n values, while for each $i, j \in V, i \neq j$, the term $\max_{S \subseteq V: S \not\ni i, j} (1 - e^{\beta \Delta_j \Delta_i f(S)})$ involves an optimization problem over 2^{n-2} possibilities. As we will see in the next section, however, it is often the case that we can compute this term exactly, or that we can easily produce upper bounds for it.

4. Applications to log submodular point processes

In this section we apply Lemma 8 to a few canonical examples of submodular functions. First, in Section 4.1 we discuss the elementary behavior of the Gibbs sampler in trivial (i.e., i.i.d.) models defined by modular functions. This case will serve to point out another deficiency (on top of the one already highlighted in Remark 3) of the theoretical results presented in Djolonga and Krause (2014) with respect to the notion of curvature. Consecutively, we consider functions that are defined

3. As done in Djolonga and Krause (2014) and Iyer and Bilmes (2015), we can also specialize our results to *supermodular* functions (f is supermodular if $-f$ is submodular). In this case we would get $\alpha(\beta) = \max_{i \in V} e^{-\beta \Delta_i f(\emptyset)}$.

on a finite graph $G = (V, E)$, where $n = |V|$ and the edge set E describes the pair of vertices that interact, Sections 4.2, 4.3, and 4.4. The strength of the pairwise interaction is modeled by a symmetric matrix L . The nature of the assumptions that our theory requires to yield fast mixing depends on the application at hand, and they involve the structure of the matrix L . Note that for the applications discussed in Sections 4.2 and 4.3 these assumptions would be satisfied if $L_{ij} \lesssim e^{-d(i,j)}$, or if $L_{ij} = 0$ whenever $d(i, j) > r$, in the spirit of the decay of correlation properties discussed in the previous section. Finally, in Section 4.5 we consider the case of decomposable functions, that is, functions that can be represented as sums of concave functions applied to modular functions. Since we need to calculate discrete derivatives, we usually prove submodularity along the way as it is nothing but $\Delta_j \Delta_i f(S) \leq 0$ for each $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$.

4.1. Modular functions

Let V be a finite set with $n = |V|$. For each $i \in V$, let $w_i \in \mathbb{R}$ be given. For each $S \subseteq V$, let $f(S) := \sum_{i \in S} w_i$ and $f(\emptyset) := 0$. It is immediate to verify that the function f is modular as $\Delta_j \Delta_i f(S) = 0$ for each $i, j \in V, i \neq j, S \subseteq V$. Clearly, in this case condition (13) is satisfied with $\gamma = 0$, and Lemma 8 implies that for each $m \geq 1$ we have $|\rho T_s^m h - \mu h| = 0$, and $|\rho T_r^m h - \mu h| = 0$. In particular, these results hold even for $m = 1$, which means that the Gibbs samplers sample *exactly* from the distribution μ in a single sweep.

We now slightly tweak the trivial example just considered to compare the theoretical guarantees provided in Lemma 8 for the Gibbs sampler against the guarantees provided in [Djolonga and Krause \(2014\)](#) for the variational methods they proposed, as a function of the *curvature* $c(f)$ of f , which, for monotone submodular functions with $f(\emptyset) \geq 0$, is defined as:

$$c(f) := 1 - \min_{i \in V: f(\{i\}) \neq 0} \frac{\Delta_i f(V \setminus \{i\})}{f(\{i\})}.$$

Fix $k, k' \in V, k \neq k'$, and let g be the function defined as $g(S) := 1$ if $S \cap \{k, k'\} \neq \emptyset$ and $g(S) := 0$ if $S \cap \{k, k'\} = \emptyset$. It is easy to check that g is submodular with

$$\Delta_j \Delta_i g(S) = \begin{cases} -1 & \text{if } \{i, j\} = \{k, k'\}, \\ 0 & \text{otherwise,} \end{cases}$$

for each $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$. For each $i \in V \setminus \{k, k'\}$ let $w_i = 1$, and let $w_k = w_{k'} = 0$. Consider the submodular function $f(S) := \sum_{i \in S} w_i + g(S) = |S \setminus \{k, k'\}| + g(S)$. Then we clearly have $c(f) = 1$, as the minimum that appears in the definition of curvature is 0 (it is attained at $i = k$, for instance), so that the bounds in [Djolonga and Krause \(2014\)](#) diverge to infinity. On the other hand, in this case the results in Lemma 8 hold with $\gamma = 1 - e^{-\beta}$. In fact, $\alpha(\beta) = \max_{i \in V} e^{-\beta \Delta_i f(V \setminus \{i\})} = 1$ and $\Delta_j \Delta_i f(S) = \Delta_j \Delta_i g(S)$, from which it follows that condition (13) reads

$$\alpha(\beta) \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left(1 - e^{\beta \Delta_j \Delta_i f(S)}\right) = 1 - e^{-\beta} =: \gamma < 1.$$

4.2. Facility location

Let $V = \{1, \dots, n\}$ be a collection of facilities, and $W = \{1, \dots, m\}$ be a collection of customers. Let $L_{k\ell} \geq 0$ be the value provided to customer $k \in V$ by the facility $\ell \in W$. For each $S \subseteq V$,

let $g(S) := \sum_{k=1}^m \max_{\ell \in S} L_{k\ell}$ and $g(\emptyset) := 0$. By definition, $g(S)$ is the maximum value that the facilities in S can provide to all customers, provided that each customer chooses to be served only the facility that provides the highest value to him. We consider the function f defined as

$$f(S) := g(S) - \lambda|S|, \quad (14)$$

where $|S|$ denotes the cardinality of the set S , and $\lambda \geq 0$. Such a function is submodular and has been considered in many applications such as large-scale clustering (Mirzasoleiman et al., 2013) and recommender systems (El-Arini et al., 2009).

Corollary 9 *For the function f in (14), Lemma 8 applies if condition (13) is replaced with*

$$e^{\lambda\beta} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \left(1 - e^{-\beta \sum_{k=1}^m L_{ki} \wedge L_{kj}}\right) \leq \gamma < 1.$$

4.3. Generalized graph cut

Let V be a vertex set, and for each $i, j \in V, i \neq j$, let $L_{ij} = L_{ji} \geq 0$ be the weight associated to the undirected edge between i and j . Let $L_{ii} = 0$ for each $i \in V$. For each $S \subseteq V$, let

$$f(S) := a + b \sum_{k \in S} \sum_{\ell \in V} L_{k\ell} - c \sum_{k \in S} \sum_{\ell \in S} L_{k\ell} \quad (15)$$

and $f(\emptyset) = a$, with $a, b, c \geq 0$ (typically a is chosen so that $f(S) \geq 0$ for any $S \subseteq V$). In the case $a = 0, b = c = 1$, we recover $f(S) = f(V \setminus S) = \sum_{k \in S} \sum_{\ell \in V \setminus S} L_{k\ell}$ and $f(\emptyset) = f(V) = 0$, which is the standard graph cut function. Namely, $f(S)$ is the sum of the weights of each edge that connects a point in S with a point in $V \setminus S$. The generalized cut function f defined above is submodular, with many applications in computer vision (Jegelka et al., 2011).

Corollary 10 *For the function f in (15), Lemma 8 applies if condition (13) is replaced with*

$$e^{\beta(2c-b) \min_{i \in V} \sum_{\ell \in V \setminus \{i\}} L_{i\ell}} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \left(1 - e^{-2c\beta L_{ij}}\right) \leq \gamma < 1.$$

4.4. Determinantal point processes

Fix n and let $L \in \mathbb{R}^{n \times n}$ be a positive definite matrix. Let $V = \{1, \dots, n\}$, and for each $S \subseteq V$ let

$$f(S) := \log \det L_S \quad (16)$$

and $f(\emptyset) := 0$, where $L_S := (L_{ij})_{i, j \in S}$. Such a function is submodular and has been used in determinantal point processes for which the partition function can be computed exactly in time $O(n^3)$. Recently, several authors have considered MCMC algorithms to sample from determinant point processes, see Shah and Ghahramani (2013) and Kang (2013) for instance. However, to the best of our knowledge, no theoretical guarantees have ever been established for the MCMC⁴ algorithms being adopted. As mentioned in the introduction, the proofs in Kang (2013) are wrong (although the sampling scheme is correct, exactly matching Algorithm 4 in the case $\beta = 1$).

4. There are efficient sampling schemes that are not based on the MCMC paradigm. See Deshpande and Rademacher (2010) and discussion therein, where they explicitly pose the problem of designing MCMC samplers.

In this section we give a probabilistic interpretation of the mechanism behind the structure of the matrix L that guarantees fast mixing MCMC algorithms in the light of the theory being developed. Let L be the covariance matrix of a collection of Gaussian random variables X^1, \dots, X^n . It can be shown that, for $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$, we have $\text{Cov}(X^i, X^j | X^S) := L_{ij} - L_{i,S}L_S^{-1}L_{S,j}$. Define the conditional correlation coefficients as follows, for $i, j \in V, S \subseteq V$:

$$\rho(i, j | S) = \frac{\text{Cov}(X^i, X^j | X^S)}{\sqrt{\text{Var}(X^i | X^S)} \sqrt{\text{Var}(X^j | X^S)}},$$

and let $I(X^i; X^j | X^S)$ be the conditional mutual information of X^i and X^j given X^S . It holds:

$$I(X^i; X^j | X^S) = -\frac{1}{2} \log(1 - \rho(i, j | S)^2). \quad (17)$$

We now show how condition (13) in Lemma 8 can be stated in terms of the quantities just introduced.

Corollary 11 *For the function f in (16), Lemma 8 applies with the condition (13) that reads*

$$\max_{i \in V} \frac{1}{\text{Var}(X^i | X^{V \setminus \{i\}})^\beta} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left(1 - e^{-2\beta I(X^i; X^j | X^S)}\right) \leq \gamma < 1.$$

If $\beta = 1$, the condition is $\max_{i \in V} \frac{1}{\text{Var}(X^i | X^{V \setminus \{i\}})} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \rho(i, j | S)^2 \leq \gamma < 1$.

Practitioners who are interested in fast mixing Gibbs samplers for determinant point processes need to exploit additional structure in the matrix L to simplify the conditions in Corollary 11.

4.5. Decomposable functions

Given a finite set V , let \mathcal{S} be a collection of subsets of V that covers V , i.e., $\mathcal{S} \subseteq 2^V$ with $\bigcup_{A \in \mathcal{S}} A = V$. For each $A \in \mathcal{S}$ let $\phi_A : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a concave function. For each $S \subseteq V$, let

$$f(S) := \sum_{A \in \mathcal{S}} \phi_A(|A \cap S|). \quad (18)$$

The function f is submodular (Stobbe and Krause, 2010). For instance, note that if $\mathcal{S} = \{S \subseteq V : |S| = 2\}$ and for each $A \in \mathcal{S}$ we have $\phi_A(0) = \phi_A(2) = -J/n, \phi_A(1) = J/n$, for a given constant $J > 0$, then (2) corresponds to the distribution of an antiferromagnetic mean-field Ising model with inverse temperature β and zero external magnetic field.

Corollary 12 *For each $A \in \mathcal{S}$ let the concave function ϕ_A be twice differentiable. Assume there exist constants $c' < 0 < c$ such that $d\phi_A(x)/dx \geq c$ and $c' \leq d^2\phi_A(x)/dx^2 \leq 0$ for all $x \in [0, n]$ and $A \in \mathcal{S}$. Then, for the function f in (18), Lemma 8 applies if condition (13) is replaced with*

$$(1 - e^{c'\beta}) e^{-c\beta \min_{i \in V} |\{A \in \mathcal{S}: A \ni i\}|} \max_{i \in V} \left| \bigcup_{A \in \mathcal{S}: A \ni i} A \right| \leq \gamma < 1.$$

Acknowledgments

We would like to thank Sekhar Tatikonda for useful discussions.

References

Soren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis: Algorithms and Analysis*. Stochastic Modelling and Applied Probability. Springer, 2007.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes*. 2007.

Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, FOCS '10*, pages 329–338. IEEE Computer Society, 2010.

Josip Djolonga and Andreas Krause. From MAP to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems 27*, pages 244–252. Curran Associates, Inc., 2014.

R. L. Dobrushin and S. B. Shlosman. Constructive criterion for the uniqueness of Gibbs field. In *Statistical physics and dynamical systems (Keszeg, 1984)*, volume 10 of *Progr. Phys.*, pages 347–370. Birkhäuser Boston, Boston, MA, 1985.

R. L. Dobrušin. Definition of a system of random variables by means of conditional distributions. *Teor. Verojatnost. i Primenen.*, 15:469–497, 1970.

Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Matrix norms and rapid mixing for spin systems. *Ann. Appl. Probab.*, 19(1):71–107, 02 2009.

Khalid El-Arini, Gaurav Veda, Dafna Shahaf, and Carlos Guestrin. Turning down the noise in the blogosphere. In *KDD*, 2009.

H. Föllmer. Tail structure of Markov chains on infinite product spaces. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 50(3):273–285, 1979.

Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, second edition, 2011.

Leslie Ann Goldberg and Mark Jerrum. The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 2007.

J Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. Determinantal processes and independence. 2006.

Rishabh Iyer and Jeff Bilmes. Submodular point processes. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS-2015)*, May 2015.

Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Fast semidifferential-based submodular function optimization. *NIPS*, 2013.

Stefanie Jegelka, Hui Lin, and Jeff A Bilmes. On fast approximate submodular minimization. In *NIPS*, 2011.

Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems 26*, pages 2319–2327. Curran Associates, Inc., 2013.

Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems (to appear)*. Cambridge University Press, February 2014.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. 2012.

D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc., 2008.

Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization: Identifying representative elements in massive data. In *NIPS*, 2013.

Elchanan Mossel and Allan Sly. Exact thresholds for Ising–Gibbs samplers on general graphs. *Ann. Probab.*, 41(1):294–328, 01 2013.

Patrick Rebeschini and Ramon van Handel. Comparison theorems for Gibbs measures. *Journal of Statistical Physics*, 157(2):234–281, 2014.

Amar Shah and Zoubin Ghahramani. Determinantal clustering processes - A nonparametric bayesian approach to kernel based semi-supervised clustering. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems 23*, pages 2208–2216. Curran Associates, Inc., 2010.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008.

Dror Weitz. Combinatorial criteria for uniqueness of Gibbs measures. *Random Structures Algorithms*, 27(4):445–475, 2005.

Appendix A. Theory, proofs

Below are the proofs of the results presented in Section 2 and Section 3.

Proof [Proof of Theorem 1] The proof is based on the Wasserstein matrix approach, which is a standard tool in the analysis of high-dimensional Markov chains, cf. Föllmer (1979). For each two distributions ρ, ρ' on \mathbb{S}_0 , and function $g : \mathbb{S}_0 \rightarrow \mathbb{R}$, we have

$$|\rho g - \rho' g| \leq |g(0) - g(1)| |\rho(\{0\}) - \rho'(\{0\})|,$$

from which it follows that, for each $i, j \in V$,

$$\delta_j(K^{[i]} g) = \max_{x \in \mathbb{S}} |K^{[i]} g(x^{V \setminus \{j\}} 0^j) - K^{[i]} g(x^{V \setminus \{j\}} 1^j)| \leq |g(0) - g(1)| C_{ij}.$$

For each function $h : \mathbb{S} \rightarrow \mathbb{R}$, $i \in V$, $x \in \mathbb{S}$, define the function $h_x^i : z^i \in \mathbb{R} \rightarrow h_x^i(z^i) := h(z^i x^{V \setminus \{i\}})$. Then, we have

$$\begin{aligned} \delta_j(T^{[i]}h) &= \max_{x \in \mathbb{S}} |K_{x^{V \setminus \{j\}}0^j}^{[i]} h_{x^{V \setminus \{j\}}0^j}^i - K_{x^{V \setminus \{j\}}1^j}^{[i]} h_{x^{V \setminus \{j\}}1^j}^i| \\ &\leq \max_{x \in \mathbb{S}} |K_{x^{V \setminus \{j\}}0^j}^{[i]} h_{x^{V \setminus \{j\}}0^j}^i - K_{x^{V \setminus \{j\}}1^j}^{[i]} h_{x^{V \setminus \{j\}}0^j}^i| + \max_{x \in \mathbb{S}} |K_{x^{V \setminus \{j\}}1^j}^{[i]} (h_{x^{V \setminus \{j\}}0^j}^i - h_{x^{V \setminus \{j\}}1^j}^i)| \\ &\leq \max_{y \in \mathbb{S}} \delta_j(K^{[i]}h_y^i) + \max_{x \in \mathbb{S}} \max_{z \in \mathbb{S}_0} |h_{x^{V \setminus \{j\}}0^j}^i(z) - h_{x^{V \setminus \{j\}}1^j}^i(z)| \\ &\leq \delta_i(h)C_{ij} + \delta_j(h)1_{i \neq j}, \end{aligned}$$

where in the last line we applied the previous bound, and we use the notation $1_{i \neq j} := 1$ if $i \neq j$, $1_{i \neq j} := 0$ otherwise. It is convenient to rewrite the previous estimate using a matrix notation:

$$\delta_j(T^{[i]}h) \leq \sum_{k \in V} \delta_k(h)(W^{[i]})_{kj}, \quad (19)$$

where for each $i \in V$, the matrix $W^{[i]} \in \mathbb{R}^{n \times n}$ is defined as $W_{kj}^{[i]} := 1_{k=i}C_{kj} + 1_{k \neq i}1_{k=j}$. The matrix $W^{[i]}$ is a *Wasserstein matrix* for the transition kernel $T^{[i]}$.

First, we apply estimate (19) to bound the systematic scan Markov chain. Iterating this estimate, we immediately find $\delta_j(T_s h) \leq \sum_{k \in V} \delta_k(h)(W_s)_{kj}$, where $W_s := W^{[i_n]} \dots W^{[i_1]}$. Given any two distributions ρ and ρ' on \mathbb{S} , by a telescoping argument we find, for any $m \geq 0$,

$$\begin{aligned} |\rho T_s^m h - \rho' T_s^m h| &\leq \max_{x, z \in \mathbb{S}} |T_s^m h(x) - T_s^m h(z)| \leq \sum_{j \in V} \delta_j(T_s^m h) \leq \sum_{i \in V} \delta_i(h) \sum_{j \in V} (W_s^m)_{ij} \\ &\leq \|W_s^m\|_\infty \sum_{i \in V} \delta_i(h) \leq \|W_s\|_\infty^m \sum_{i \in V} \delta_i(h) \leq \gamma^m \sum_{i \in V} \delta_i(h), \end{aligned}$$

where for the last inequality we used that $\|W_s\|_\infty \leq \|C\|_\infty \leq \gamma < 1$ (this fact is proved in Corollary 24 in Dyer et al. (2009); note that the $\|\cdot\|_1$ matrix norm in the authors's notation corresponds to the $\|\cdot\|_\infty$ matrix norm in our notation).

We now apply estimate (19) to bound the random scan Markov chain. As an intermediate step, define $S := \frac{1}{n} \sum_{i \in V} T^{[i]}$. We have

$$\delta_j(Sh) \leq \frac{1}{n} \sum_{i \in V} \delta_j(T^{[i]}h) \leq \frac{1}{n} \sum_{i \in V} \sum_{k \in V} \delta_k(h)(W^{[i]})_{kj} = \sum_{i \in V} \delta_i(h)Z_{ij},$$

where $Z := \frac{1}{n}C + \frac{n-1}{n}I \in \mathbb{R}^{n \times n}$, I being the identity matrix. Iterating this result n times, we find $\delta_j(T_r h) \leq \sum_{i \in V} \delta_i(h)(W_r)_{ij}$, where $W_r = Z^n = (\frac{1}{n}C + \frac{n-1}{n}I)^n$. By the same argument presented above, we get

$$|\rho T_r^m h - \rho' T_r^m h| \leq \|W_r\|_\infty^m \sum_{i \in V} \delta_i(h) \leq \left(1 - \frac{1-\gamma}{n}\right)^{nm} \sum_{i \in V} \delta_i(h) \leq (e^{\gamma-1})^m \sum_{i \in V} \delta_i(h).$$

The proof is concluded if we take ρ' to be μ , noticing that $\mu T_s = \mu T_r = \mu$. ■

Proof [Proof of Lemma 5] Each kernel $T^{[i]}$ leaves μ invariant as a consequence of the tower property of conditional expectations as, for each function $h : \mathbb{S} \rightarrow \mathbb{R}$ we have

$$\mu T^{[i]} h = \mathbf{E}[T^{[i]} h(X)] = \mathbf{E}[\mathbf{E}[h(X)|X^{V \setminus \{v\}}]] = \mathbf{E}[h(X)] = \mu h.$$

For $i, j \in V$, $i \neq j$, $x \in \mathbb{S}$, we have

$$\begin{aligned} |K_{x^{V \setminus \{j\}} 0^j}^{[i]}(\{0\}) - K_{x^{V \setminus \{j\}} 1^j}^{[i]}(\{0\})| &= \left| \frac{1}{1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 0^j))}} - \frac{1}{1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 1^j))}} \right| \\ &= \frac{|e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 0^j))} - e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 1^j))}|}{(1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 0^j))})(1 + e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 1^j))})}. \end{aligned}$$

The proof is concluded taking the maximum over $x \in \mathbb{S}$ on both hand sides. \blacksquare

Proof [Proof of Lemma 6] The fact that each kernel $T^{[i]}$ leaves μ invariant follows as a consequence of the so-called *detailed balance equation*, which holds for each $x \in \mathbb{S}$, $y^i \in \mathbb{S}_0$ and is immediately verified (once noticed that $\alpha^i(x) = 1 \wedge \frac{\mu(x^{V \setminus \{i\}}(1-x^i)^i)}{\mu(x)}$):

$$\mu(x) K^{[i]}(x, y^i) = \mu(x^{V \setminus \{i\}} y^i) K^{[i]}(x^{V \setminus \{i\}} y^i, x^i).$$

In fact, this equation yields, for each function $h : \mathbb{S} \rightarrow \mathbb{R}$,

$$\begin{aligned} \mu T^{[i]} h &= \sum_{x \in \mathbb{S}} \sum_{y^i \in \mathbb{S}_0} \mu(x) K^{[i]}(x, y^i) h(x^{V \setminus \{i\}} y^i) \\ &= \sum_{x \in \mathbb{S}} \sum_{y^i \in \mathbb{S}_0} \mu(x^{V \setminus \{i\}} y^i) K^{[i]}(x^{V \setminus \{i\}} y^i, x^i) h(x^{V \setminus \{i\}} y^i), \end{aligned}$$

and the right-hand side of this expression is equal to μh , as clearly $\sum_{x^i \in \mathbb{S}_0} K^{[i]}(x^{V \setminus \{i\}} y^i, x^i) = 1$. For $i \in V$, $x \in \mathbb{S}$, we have

$$|K_{x^{V \setminus \{i\}} 0^i}^{[i]}(\{0\}) - K_{x^{V \setminus \{i\}} 1^i}^{[i]}(\{0\})| = |1 - \alpha^i(x^{V \setminus \{i\}} 0^i) - \alpha^i(x^{V \setminus \{i\}} 1^i)| = e^{-\beta |\Delta_i f(S(x^{V \setminus \{i\}} 0^i))|}.$$

For $i, j \in V$, $i \neq j$, $x \in \mathbb{S}$, we have

$$\begin{aligned} |K_{x^{V \setminus \{i,j\}} 0^i 0^j}^{[i]}(\{0\}) - K_{x^{V \setminus \{i,j\}} 0^i 1^j}^{[i]}(\{0\})| &= |1 \wedge e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 0^j))} - 1 \wedge e^{\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 0^i 1^j))}|, \\ |K_{x^{V \setminus \{i,j\}} 1^i 0^j}^{[i]}(\{0\}) - K_{x^{V \setminus \{i,j\}} 1^i 1^j}^{[i]}(\{0\})| &= |1 \wedge e^{-\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 1^i 0^j))} - 1 \wedge e^{-\beta \Delta_i f(S(x^{V \setminus \{i,j\}} 1^i 1^j))}|. \end{aligned}$$

Hence, the Dobrushin coefficients (7) read

$$\begin{aligned} \tilde{C}_{ij} &= \max_{x \in \mathbb{S}} |K_{x^{V \setminus \{j\}} 0^j}^{[i]}(\{0\}) - K_{x^{V \setminus \{j\}} 1^j}^{[i]}(\{0\})| = \\ &\begin{cases} \max_{S \subseteq V: S \not\ni i} e^{-\beta |\Delta_i f(S)|} & \text{if } i = j, \\ \max_{S \subseteq V: S \not\ni i, j} \left| 1 \wedge e^{\beta \Delta_i f(S)} - 1 \wedge e^{\beta \Delta_i f(S \cup \{j\})} \right| \vee \left| 1 \wedge e^{-\beta \Delta_i f(S)} - 1 \wedge e^{-\beta \Delta_i f(S \cup \{j\})} \right| & \text{if } i \neq j. \end{cases} \end{aligned}$$

\blacksquare

Below are the algorithms considered in Section 3.

Algorithm 3 Random scan Gibbs sampler

Sample $S \subseteq V$ from the distribution ρ ;
for $k = 1, \dots, nm$ **do**
 | Sample $i \in V$ uniformly;
 | Draw $C \in \{0, 1\}$ with $\mathbf{P}(C = 0) = \frac{1}{1+e^{\Delta_i f(S \setminus \{i\})}}$;
 | If $C = 0$ then set $S \leftarrow S \setminus \{i\}$, else set $S \leftarrow S \cup \{i\}$;
end
Output: $S \subseteq V$ that is distributed according to ρT_r^m .

Algorithm 4 Random scan Metropolis-Hastings

Sample $S \subseteq V$ from the distribution ρ ;
for $k = 1, \dots, nm$ **do**
 | Sample $i \in V$ uniformly;
 | **if** $i \in S$ **then**
 | | draw $C \in \{0, 1\}$ with $\mathbf{P}(C = 0) = 1 \wedge e^{-\beta \Delta_i f(S \setminus \{i\})}$; if $C = 0$, then set $S \leftarrow S \setminus \{i\}$;
 | **else**
 | | draw $C \in \{0, 1\}$ with $\mathbf{P}(C = 1) = 1 \wedge e^{\beta \Delta_i f(S \setminus \{i\})}$; if $C = 1$, then set $S \leftarrow S \cup \{i\}$;
 | **end**
end
Output: $S \subseteq V$ that is distributed according to ρT_r^m .

Proof [Proof of Lemma 7] The first statement is trivial as $C_{ii} = 0$ and $\tilde{C}_{ii} \geq 0$ by definition. To prove the second statement, we show that the following holds for each pair of real numbers a, b with $b \leq a$:

$$\frac{1}{4}h(a, b) \leq \frac{e^a - e^b}{(1 + e^a)(1 + e^b)} \leq h(a, b),$$

where $h(a, b) := (1 \wedge e^a - 1 \wedge e^b) \vee (1 \wedge e^{-b} - 1 \wedge e^{-a})$. We distinguish three cases.

If $0 \leq b \leq a$, then $h(a, b) = e^{-b} - e^{-a}$, and we have

$$\frac{1}{4}h(a, b) = \frac{e^{-b} - e^{-a}}{4} = \frac{e^a - e^b}{4e^a e^b} \leq \frac{e^a - e^b}{(1 + e^a)(1 + e^b)} \leq \frac{e^a - e^b}{e^a e^b} = e^{-b} - e^{-a} = h(a, b).$$

If $b \leq 0 \leq a$, then $h(a, b) = (1 - e^b) \vee (1 - e^{-a}) = 1 - e^{(-a) \wedge b}$. As

$$\frac{1 - e^{2((-a) \wedge b)}}{2} \leq \frac{1 - e^{2((-a) \wedge b)}}{1 + e^{(-a) \wedge b - (-a) \vee b}} \leq 1 - e^{(-a) \wedge b + (-a) \vee b} = 1 - e^{-a+b},$$

where for the second inequality we used that $\frac{1 - e^{2x}}{1 + e^{x-y}} \leq 1 - e^{x+y}$ for $x \leq y \leq 0$, it follows that

$$\frac{1}{4}h(a, b) \leq \frac{1}{2} \frac{1 - e^{(-a) \wedge b}}{1 + e^{(-a) \vee b}} = \frac{1}{2} \frac{1 - e^{2((-a) \wedge b)}}{(1 + e^{-a})(1 + e^b)} \leq \frac{1 - e^{-a+b}}{(1 + e^{-a})(1 + e^b)} = \frac{e^a - e^b}{(1 + e^a)(1 + e^b)}.$$

Moreover,

$$\frac{e^a - e^b}{(1 + e^a)(1 + e^b)} = \frac{1 - e^{-a+b}}{(1 + e^{-a})(1 + e^b)} \leq \frac{1 - e^{2((-a) \wedge b)}}{(1 + e^{(-a) \wedge b})(1 + e^{(-a) \vee b})} = \frac{1 - e^{(-a) \wedge b}}{1 + e^{(-a) \vee b}} \leq h(a, b).$$

Finally, if $b \leq a \leq 0$, then $h(a, b) = e^a - e^b$, and we have

$$\frac{1}{4}h(a, b) = \frac{e^a - e^b}{4} \leq \frac{e^a - e^b}{(1 + e^a)(1 + e^b)} \leq e^a - e^b = h(a, b).$$

■

Appendix B. Applications, proofs

Below are the proofs of the results presented in Section 4.

Proof [Proof of Lemma 8] For each $i, j \in V$, define

$$R_{ij} := \begin{cases} 0 & \text{if } i = j, \\ \alpha(\beta) \max_{S \subseteq V: S \not\ni i, j} \left| 1 - e^{\beta \Delta_j \Delta_i f(S)} \right| & \text{if } i \neq j, \end{cases}$$

with $\alpha(\beta) := \max_{i \in V} \max_{S \subseteq V \setminus \{i\}} e^{-\beta \Delta_i f(S)}$. The first part of the Lemma follows immediately from Theorem 1 and Remark 2, once we prove that $C \leq R$, element-wise, where C is the Dobrushin matrix defined in Lemma 5. In fact, condition (8) yields

$$\|R\|_\infty = \max_{i \in V} \sum_{j \in V} R_{ij} \leq \alpha(\beta) \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left| 1 - e^{\beta \Delta_j \Delta_i f(S)} \right| \leq \gamma < 1,$$

which corresponds to (12). For each $i \in V$ we clearly have $C_{ii} = R_{ii} = 0$. Henceforth, fix $i, j \in V$, $i \neq j$. As

$$\frac{|e^{\beta \Delta_i f(S)} - e^{\beta \Delta_i f(S \cup \{j\})}|}{(1 + e^{\beta \Delta_i f(S)})(1 + e^{\beta \Delta_i f(S \cup \{j\})})} \leq |e^{-\beta \Delta_i f(S \cup \{j\})} - e^{-\beta \Delta_i f(S)}| = e^{-\beta \Delta_i f(S \cup \{j\})} \left| 1 - e^{\beta \Delta_j \Delta_i f(S)} \right|,$$

taking the maximum over $S \subseteq V, S \not\ni i, j$, on both hand sides immediately yields $C_{ij} \leq R_{ij}$. The proof of the Lemma is concluded once noticed that f being submodular means that $\Delta_j \Delta_i f(S) \leq 0$ for each $i, j \in V, i \neq j, S \subseteq V, S \not\ni i, j$, and it implies $\Delta_i f(V \setminus \{i\}) \leq \Delta_i f(S)$ for each $i \in V, S \subseteq V, S \not\ni i$, so that $\alpha(\beta) = \max_{i \in V} e^{-\beta \Delta_i f(V \setminus \{i\})}$. ■

Proof [Proof of Corollary 9] For each $i \in V, S \subseteq V, i \notin S$, we have $\Delta_i g(S) = \sum_{k=1}^m (L_{ki} - \max_{\ell \in S} L_{k\ell}) \vee 0$, and, clearly, $\Delta_i f(S) = \Delta_i g(S) - \lambda$. The function g is submodular as, for each $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$, we have

$$\Delta_i g(S \cup \{j\}) = \sum_{k=1}^m (L_{ki} - \max_{\ell \in S \cup \{j\}} L_{k\ell}) \vee 0 \leq \sum_{k=1}^m (L_{ki} - \max_{\ell \in S} L_{k\ell}) \vee 0 = \Delta_i g(S),$$

so also the function f is submodular, as $\Delta_j \Delta_i f(S) = \Delta_j \Delta_i g(S) \leq 0$. Note that for each $i, j \in V$, $i \neq j$, $S \subseteq V$, $i, j \notin S$, we can write

$$\begin{aligned}\Delta_j \Delta_i f(S) &= \sum_{k=1}^m (L_{ki} - \max_{\ell \in S \cup \{j\}} L_{k\ell}) \vee 0 - \sum_{k=1}^m (L_{ki} - \max_{\ell \in S} L_{k\ell}) \vee 0 \\ &= \sum_{k=1}^m \left(L_{ki} \wedge \max_{\ell \in S} L_{k\ell} - L_{ki} \wedge \max_{\ell \in S \cup \{j\}} L_{k\ell} \right) \\ &= \sum_{k=1}^m \left(L_{ki} \wedge \max_{\ell \in S} L_{k\ell} - L_{ki} \wedge (L_{kj} \vee \max_{\ell \in S} L_{k\ell}) \right) \\ &= \sum_{k=1}^m \left\{ ((\max_{\ell \in S} L_{k\ell} - L_{ki}) \vee 0) 1_{L_{ki} < L_{kj}} + ((\max_{\ell \in S} L_{k\ell} - L_{kj}) \vee 0) 1_{L_{ki} \geq L_{kj}} \right\},\end{aligned}$$

where we used, in order, the following two equalities holding for any real numbers x, y, z :

$$(x - y) \vee 0 = x - y \wedge x, \quad x \wedge z - x \wedge (y \vee z) = ((z - x) \vee 0) 1_{x < y} + ((z - y) \vee 0) 1_{x \geq y}.$$

From the expression above it is clear that for each $i, j \in V$, $i \neq j$, we have $\Delta_j \Delta_i f(S) \leq \Delta_j \Delta_i f(S')$ if $S \subseteq S' \subseteq V$, from which it follows that $\min_{S \subseteq V: S \not\ni i, j} \Delta_j \Delta_i f(S) = \Delta_j \Delta_i f(\emptyset) = -\sum_{k=1}^m L_{ki} \wedge L_{kj}$. As $\min_{i \in V} \Delta_i f(V \setminus \{i\}) \geq -\lambda$, the left hand side of (13) is upper bounded by

$$e^{\lambda\beta} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \left(1 - e^{-\beta \sum_{k=1}^m L_{ki} \wedge L_{kj}} \right).$$

■

Proof [Proof of Corollary 10] It is easy to check that for each $i, j \in V$, $i \neq j$, and $S \subseteq V$ so that $i, j \notin S$ we have $\Delta_i f(S) = b \sum_{\ell \in V} L_{i\ell} - 2c \sum_{\ell \in S} L_{i\ell}$ and $\Delta_j \Delta_i f(S) = -2c L_{ij} \leq 0$, from which it follows that f is submodular. As $\min_{i \in V} \Delta_i f(V \setminus \{i\}) = (b - 2c) \min_{i \in V} \sum_{\ell \in V \setminus \{i\}} L_{i\ell}$, the left hand side of (13) is upper bounded by

$$e^{\beta(2c-b) \min_{i \in V} \sum_{\ell \in V \setminus \{i\}} L_{i\ell}} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \left(1 - e^{-2c\beta L_{ij}} \right).$$

■

Proof [Proof of Corollary 11] For $i \notin S \subseteq V$ we have

$$\det L_{S \cup \{i\}} = (\det L_S)(L_{ii} - L_{i,S} L_S^{-1} L_{S,i}),$$

where $L_{i,S} := (L_{ij})_{j \in S} \in \mathbb{R}^{1 \times |S|}$ and $L_{S,i} := (L_{ji})_{j \in S} \in \mathbb{R}^{|S| \times 1}$. It follows that

$$f(S \cup \{i\}) = \log \det L_{S \cup \{i\}} = f(S) + \log(L_{ii} - L_{i,S} L_S^{-1} L_{S,i}),$$

and the marginal gain reads $\Delta_i f(S) = \log(L_{ii} - L_{i,S}L_S^{-1}L_{S,i}) = \log \text{Var}(X^i|X^S)$. Analogously, if $i \notin S, i \neq j$, we find $\Delta_i f(S \cup \{j\}) = \log(L_{ii} - L_{i,S \cup \{j\}}L_{S \cup \{j\}}^{-1}L_{S \cup \{j\},i})$. Recall that

$$L_{S \cup \{j\}}^{-1} = \begin{pmatrix} B & C \\ C^T & \frac{1}{d} \end{pmatrix},$$

with $d := L_{jj} - L_{j,S}L_S^{-1}L_{S,j}$, $B := L_S^{-1} + \frac{1}{d}L_S^{-1}L_{S,j}L_{j,S}L_S^{-1}$, and $C := -\frac{1}{d}L_S^{-1}L_{S,j}$. We get

$$L_{i,S \cup \{j\}}L_{S \cup \{j\}}^{-1}L_{S \cup \{j\},i} = (L_{i,S}, L_{ij}) \begin{pmatrix} B & C \\ C^T & \frac{1}{d} \end{pmatrix} \begin{pmatrix} L_{S,i} \\ L_{ji} \end{pmatrix} = L_{i,S}BL_{S,i} + 2L_{ji}L_{i,S}C + \frac{L_{ij}^2}{d},$$

where we used that $L_{ij} = L_{ji}$, and that $L_{ij}C^T L_{S,i} = L_{ij}(L_{S,i}^T C)^T = L_{ji}(L_{i,S}C)^T = L_{ji}L_{i,S}C$. Consequently,

$$\begin{aligned} L_{i,S \cup \{j\}}L_{S \cup \{j\}}^{-1}L_{S \cup \{j\},i} &= L_{i,S} \left(L_S^{-1} + \frac{1}{d}L_S^{-1}L_{S,j}L_{j,S}L_S^{-1} \right) L_{S,i} + 2L_{ji}L_{i,S} \left(-\frac{1}{d}L_S^{-1}L_{S,j} \right) + \frac{L_{ij}^2}{d} \\ &= L_{i,S}L_S^{-1}L_{S,i} + \frac{(L_{ij} - L_{i,S}L_S^{-1}L_{S,j})^2}{L_{jj} - L_{j,S}L_S^{-1}L_{S,j}}, \end{aligned}$$

where we used that $L_{j,S}L_S^{-1}L_{S,i} = (L_{j,S}L_S^{-1}L_{S,i})^T = L_{S,i}^T(L_S^{-1})^T L_{j,S}^T = L_{i,S}(L_S^T)^{-1}L_{S,j} = L_{i,S}L_S^{-1}L_{S,j}$. Therefore, we find

$$\Delta_j \Delta_i f(S) = \log \frac{L_{ii} - L_{i,S \cup \{j\}}L_{S \cup \{j\}}^{-1}L_{S \cup \{j\},i}}{L_{ii} - L_{i,S}L_S^{-1}L_{S,i}} = \log(1 - \rho(i, j|S)^2).$$

As $0 \leq \rho(i, j|S) \leq 1$ for each $i, j \in V, S \subseteq V$, it follows that $\Delta_j \Delta_i f(S) \leq 0$ and f is submodular. From (17) we have $\Delta_j \Delta_i f(S) = -2I(X^i; X^j|X^S)$, and the left hand side of (13) is equal to

$$\max_{i \in V} \frac{1}{\text{Var}(X^i|X^{V \setminus \{i\}})^\beta} \max_{i \in V} \sum_{j \in V \setminus \{i\}} \max_{S \subseteq V: S \not\ni i, j} \left(1 - e^{-2\beta I(X^i; X^j|X^S)} \right).$$

The case $\beta = 1$ follows immediately using (17). ■

Proof [Proof of Corollary 12] First note that since $d\phi_A(x)/dx \geq c$ for each $x \in [0, n]$ it follows that the discrete differences are all bounded, i.e., $\phi_A(x+1) - \phi_A(x) \geq c$ for each $x \in \{0, \dots, n-1\}$ and $A \in \mathcal{S}$. Similarly, since $d^2\phi_A(x)/dx^2 \geq c'$ for each $x \in [0, n]$ we have $\phi_A(x+2) - 2\phi_A(x+1) + \phi_A(x) \geq c'$ for each $x \in \{0, \dots, n-2\}$ and $A \in \mathcal{S}$. Now, for each $i \in V, S \subseteq V, i \notin S$ we have

$$\Delta_i f(S) = \sum_{A \in \mathcal{S}: A \ni i} \{\phi_A(|A \cap (S \cup \{i\})|) - \phi_A(|A \cap S|)\},$$

and each $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$ we have

$$\begin{aligned} \Delta_j \Delta_i f(S) &= \sum_{A \in \mathcal{S}: A \ni i, j} \{\phi_A(|A \cap (S \cup \{i, j\})|) - \phi_A(|A \cap (S \cup \{j\})|) - \phi_A(|A \cap (S \cup \{i\})|) + \phi_A(|A \cap S|)\} \\ &= \sum_{A \in \mathcal{S}: A \ni i, j} \{\phi_A(|A \cap S| + 2) - 2\phi_A(|A \cap S| + 1) + \phi_A(|A \cap S|)\} \leq 0, \end{aligned}$$

where the inequality comes from the concavity of each ϕ_A , as $\phi_A(x+2) - 2\phi_A(x+1) + \phi_A(x) \leq 0$ for $x \geq 0$. As $\Delta_j \Delta_i f(S) \leq 0$, f is submodular. In particular, note that if i, j are such that there is no $A \in \mathcal{S}$ that satisfies $A \ni i, j$, then the previous expression yields $\Delta_j \Delta_i f(S) = 0$. As for each $i \in V$ we have $\Delta_i f(V \setminus \{i\}) = \sum_{A \in \mathcal{S}: A \ni i} \{\phi_A(|A|) - \phi_A(|A| - 1)\} \geq c|\{A \in \mathcal{S} : A \ni i\}|$, and for each $i, j \in V, i \neq j, S \subseteq V, i, j \notin S$ we have $\Delta_j \Delta_i f(S) \geq c'$, then the left hand side of (13) is upper bounded by

$$(1 - e^{c'\beta})e^{-c\beta \min_{i \in V} |\{A \in \mathcal{S} : A \ni i\}|} \max_{i \in V} |\{j \in V \setminus \{i\} : i, j \in A \text{ for some } A \in \mathcal{S}\}|$$

$$\leq (1 - e^{c'\beta})e^{-c\beta \min_{i \in V} |\{A \in \mathcal{S} : A \ni i\}|} \max_{i \in V} \left| \bigcup_{A \in \mathcal{S} : A \ni i} A \right|.$$

■